日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年１１月１１日

出願番号
Application Number: 特願２００２－３２６５９８

[ST.10/C]: [ＪＰ２００２－３２６５９８]

出願人
Applicant(s): 株式会社日立製作所

２００３年　６月１３日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０４６４９２

【書類名】　特許願

【整理番号】　HI020566

【提出日】　平成14年11月11日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F　3/06

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　高田　豊

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　小林　直孝

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　小笠原　裕

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社日立製作所

【代理人】

【識別番号】　100071283

【弁理士】

【氏名又は名称】　一色　健輔

【選任した代理人】

【識別番号】　100084906

【弁理士】

【氏名又は名称】　原島　典孝

【選任した代理人】

【識別番号】　100098523

【弁理士】

【氏名又は名称】　黒川　恵

【選任した代理人】

【識別番号】　100112748

【弁理士】

【氏名又は名称】　吉田　浩二

【選任した代理人】

【識別番号】　100110009

【弁理士】

【氏名又は名称】　青木　康

【手数料の表示】

【予納台帳番号】　011785

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　ディスク制御装置およびディスク制御装置の制御方法

【特許請求の範囲】

【請求項１】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段と、

前記ネットワーク制御部が複数のアドレスが設定されている１の前記コマンドをディスク制御部に送信する手段と、

前記ディスク制御部が前記コマンドを受信してこのコマンドに設定されている前記各アドレスに対応するデータ入出力をディスクドライブに対して行う手段と、

を備えることを特徴とするディスク制御装置。

【請求項２】　請求項１に記載のディスク制御装置において、

前記ネットワーク制御部ではファイルシステムが動作し、

前記データ入出力要求は、ディスクドライブに入出力されるデータをファイル名により指定するものであり、

前記ネットワーク制御部は、前記データ入出力要求に設定されているファイル名に対応するデータのディスクドライブ上の記憶位置に対応するアドレスを生成し、このアドレスを前記コマンドに設定する手段を備えること、

を特徴とするディスク制御装置。

【請求項３】　請求項１に記載のディスク制御装置において、

前記アドレスは前記ディスクドライブの記憶領域に区画設定された論理的な記憶領域におけるデータの記憶位置を指定する論理的なアドレスであること、

を特徴とするディスク制御装置。

【請求項４】　請求項１に記載のディスク制御装置において、前記内部バス

はＰＣＩバスであること、

を特徴とするディスク制御装置。

【請求項５】　請求項１に記載のディスク制御装置において、前記ネットワーク制御部はネットワークプロトコルに従って前記外部装置と通信する手段を備えること、

を特徴とするディスク制御装置。

【請求項６】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段を備え、

前記回路基板に前記ネットワーク制御部および前記ディスク制御部が共通してアクセス可能なメモリを有し、

前記ネットワーク制御部および前記ディスク制御部が設定されたタイミングで前記メモリに自身の動作状態を示す動作状態情報を更新する手段を備え、

前記動作状態情報に基づいて前記ネットワーク制御部もしくは前記ディスク制御部に障害が発生したことを検知する手段を備えること、

を特徴とするディスク制御装置。

【請求項７】　請求項６に記載のディスク制御装置において、

前記ネットワーク制御部は、前記コマンドを前記ディスク制御部に送信するに際し前記コマンドの送信先となる前記ディスク制御部の動作状態を前記動作状態情報から取得し、取得した動作状態に応じて前記コマンドを前記ディスク制御部に送信するかどうかを決定すること、

を特徴とするディスク制御装置。

【請求項８】　請求項６に記載のディスク制御装置において、

前記ネットワーク制御部は、前記ディスク制御部に送信したコマンドについての受信通知を取得できない場合に前記動作状態情報に基づいて送信先の前記ディ

スク制御部の動作状態を調査し、その調査の結果に応じてそのコマンドを再び前記ディスク制御部に送信するかどうかを判断すること、

を特徴とするディスク制御装置。

【請求項９】　請求項６に記載のディスク制御装置において、

前記ネットワーク制御部は、前記ディスク制御部に送信したコマンドについての受信通知を取得できない場合に前記動作状態情報に基づいて送信先の前記ディスク制御部の動作状態を調査し、前記ディスク制御装置が正常に動作していないと判断した場合には、他の前記ディスク制御部に前記コマンドを送信すること、

を特徴とするディスク制御装置。

【請求項１０】　請求項６に記載のディスク制御装置において、

前記障害を検知した場合にその旨を通知するユーザインタフェースを備えることを特徴とするディスク制御装置。

【請求項１１】　請求項６に記載のディスク制御装置において、

前記障害を検知した場合に、障害が発生している前記ネットワーク制御部もしくは前記ネットワーク制御部に、再起動を要求する信号を送信すること、

を特徴とするディスク制御装置。

【請求項１２】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段を備え、

前記ディスク制御部がバックアップ装置との接続手段を備え、

前記ネットワーク制御部が前記ディスクドライブに記憶しているデータについてのバックアップ要求を外部装置から受信して前記ディスク制御部にバックアップコマンドを送信する手段を備え、

前記ディスク制御部はが前記バックアップコマンドを受信すると前記ディスクドライブのデータについてのバックアップ指示をバックアップ装置に送信する手

段を備えること、

を特徴とするディスク制御装置。

【請求項１３】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うディスク制御装置において、

互いに通信可能に接続された複数の前記回路基板を備え、

前記回路基板間でハートビートメッセージを交換することで１の前記回路基板に障害が生じたことを他の回路基板が検知する手段を備え、

前記回路基板が他の前記回路基板についての障害を検知した場合に、障害となっている前記回路基板が行っている処理をその回路基板とは異なる他の回路基板に代行させる手段を備えること、

を特徴とするディスク制御装置。

【請求項１４】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段を備えることを特徴とするディスク制御装置。

【請求項１５】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、前記ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うディスク制御装置を、

前記ネットワーク制御部が、複数のアドレスが設定されている１の前記コマン

ドをディスク制御部に送信し、

前記ディスク制御部が、前記コマンドを受信してこのコマンドに設定されている各アドレスに対応するデータ入出力をディスクドライブに対して行うようにすること、

を特徴とするディスク制御装置の制御方法。

【請求項１６】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、前記ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、前記回路基板に前記ネットワーク制御部および前記ディスク制御部が共通してアクセス可能なメモリを有するディスク制御装置を、

前記ネットワーク制御部および前記ディスク制御部が、設定されたタイミングで前記メモリに自身の動作状態を示す動作状態情報を更新し、

前記動作状態情報に基づいて前記ネットワーク制御部もしくは前記ディスク制御部に障害が発生したことを検知するようにすること、

を特徴とするディスク制御装置の制御方法。

【請求項１７】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、ディスク制御部はバックアップ装置との接続手段を備えるディスク制御装置を、

前記ネットワーク制御部が、前記ディスクドライブに記憶しているデータについてのバックアップ要求を外部装置から受信して前記ディスク制御部にバックアップコマンドを送信し、

前記ディスク制御部が、前記バックアップコマンドを受信すると前記ディスク

ドライブのデータについてのバックアップ指示をバックアップ装置に送信するようにすること、

を特徴とするディスク制御装置の制御方法。

【請求項１８】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、互いに通信可能に接続された複数の前記回路基板を備えるディスク制御装置を、

前記回路基板間でハートビートメッセージを交換することで１の前記回路基板に障害が生じたことを他の回路基板が検知し、

前記回路基板が他の前記回路基板についての障害を検知した場合に、障害となっている前記回路基板が行っている処理をその回路基板とは異なる他の回路基板に代行させるようにすること、

を特徴とするディスク制御装置の制御方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、ディスク制御装置およびディスク制御装置の制御方法に関する。

【０００２】

【従来の技術】

ＳＡＮ（Storage Area Network）を用いて構成されるストレージシステムの典型的な構成を図１２に示す。ディスク制御装置６０が提供する記憶領域にアクセスしようとする情報処理装置であるホストコンピュータ２０は、ＬＡＮ（Local Area Network）５０を通じてサーバコンピュータ７０にアクセスする。サーバコンピュータ７０とディスク制御装置６０とはファイバチャネル８０により接続されている。

【０００３】

サーバコンピュータ７０ではファイルシステム７５が動作している。ホストコンピュータ２０からサーバコンピュータ７０へは、ファイル指定によるデータ入出力要求が送信される。ファイルシステム７５はこのデータ入出力要求に基づいてＳＣＳＩ規格などに従ったＩ／Ｏコマンドを生成し、生成したＩ／Ｏコマンドをファイバチャネル８０を通じてディスク制御装置６０に送信する。ディスク制御装置６０はファイバチャネル８０を通じて送られてくるＩ／Ｏコマンドに従ってディスクドライブ９０に対してデータ入出力を行う。ディスク制御装置６０は、ディスクドライブ９０から読み出したデータや処理完了報告などをサーバコンピュータ７０に送信し、さらにサーバコンピュータ７０からホストコンピュータ２０に読み出したデータや処理完了報告が通知される。

【０００４】

【特許文献１】

特開平８－３３５１４４号公報

【０００５】

【発明が解決しようとする課題】

ところで、ファイバチャネル８０による通信は、ファイバチャネル規格やＳＣＳＩ規格に従って行われるが、これらの規格は、時としてサーバコンピュータ７０やディスク制御装置のハードウエアが有する潜在的な性能や能力を引き出すことの妨げとなる。またファイバチャネル８０で通信を行うためには、ディスク制御装置６０やサーバコンピュータ７０内で稼働するＣＰＵやメモリ間で行われる内部バスを通じた通信とファイバチャネル８０を用いた通信との間でのプロトコル変換のための回路が必要であり、これにより装置の複雑化やコスト増を招く。またこの変換にかかる処理のオーバーヘッドはストレージシステムの性能を低下させる要因となる。

【０００６】

本発明は、このような事情に鑑みてなされたもので、ディスク制御装置およびディスク制御装置の制御方法を提供することを目的とする。

【０００７】

【課題を解決するための手段】

この目的を達成するための主たる発明にかかるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段と、

前記ネットワーク制御部が複数のアドレスが設定されている１の前記コマンドをディスク制御部に送信する手段と、

前記ディスク制御部が前記コマンドを受信してこのコマンドに設定されている前記各アドレスに対応するデータ入出力をディスクドライブに対して行う手段と、

を備えることとする。

なお、本発明の他の特徴については、本明細書及び添付図面の記述により明らかにする。

【０００８】

【発明の実施の形態】

＜開示の概要＞

以下の開示により、少なくともつぎのことが明らかにされる。

前記の発明において、ネットワーク制御部は、ＬＡＮを通じてホストコンピュータ２０から所定のネットワークプロトコルに従って送られてくるデータ入出力要求を受信するという、従来のストレージシステムにおけるサーバコンピュータが行っていた機能を提供する。またネットワーク制御部は、ウイルスチェックやＳＮＭＰ（Simple Network Management Protocol）、クラスタ管理、ファイルに対するアクセス制限や時刻管理などのシステム管理の機能などを備えていることもある。また、ディスク制御部は、ＳＡＮを通じてサーバコンピュータからコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うという、従来のストレージシステムにおけるディスク制御装置の機能を提供する。なお、データ入出力という場合には、ディスクドライブへのデータの書き

込みもしくはディスクドライブからの読み出しのうち少なくともいずれかが含まれる。前記コマンドは、例えば、後述するＩ／Ｏコマンドが対応する。

【０００９】

このディスク制御装置では、ネットワーク制御部とディスク制御部とが同一回路基板上に形成され、両者がこの回路基板に設けられたＰＣＩバス（Peripheral Component Interconnect Bus）などの内部バスで接続されていることで、ファイバチャネルによる通信の場合のようにプロトコルの制限に束縛されることなくネットワーク制御部とディスク制御部とは自由度の高い通信を行うことが可能である。

【００１０】

また、前記のようにネットワーク制御部が複数のアドレスが設定されている１の前記コマンドをディスク制御部に送信し、ディスク制御部が前記コマンドを受信してこのコマンドに設定されている各アドレスに対応するデータ入出力をディスクドライブに対して行うようにすることが可能である。

【００１１】

従って、従来のファイバチャネルによる通信では、あるデータ入出力要求を受信したサーバコンピュータがディスク制御装置に複数のコマンドを送信する必要があった処理を１のコマンドの送信により行うことができ、これにより通信にかかるオーバーヘッドが軽減されてストレージシステムの性能向上が図られる。

【００１２】

また、ネットワーク制御部とディスク制御部とが回路基板に設けられた内部バスを通じて接続することで、前述のようなファイバチャネルの通信と内部バスの通信との間の変換のための回路も必要なく、ストレージシステムの生産性の向上やコスト低減も図られる。

【００１３】

また、ディスク制御部がディスクドライブへのデータ入出力に関する処理について並行処理が可能な場合には、１のコマンドで複数のアドレスに対応するデータ入出力に関する処理を並行して行うことが可能となり、これによりストレージシステムの性能向上が図られる。また、ネットワーク制御部とディスク制御部と

を同一回路基板に形成することで、これらが異なる回路基板で構成される場合に比べて製造工程の簡素化を図ることができる。

【００１４】

また、ネットワーク制御部は、ディスクドライブに入出力されるデータをファイル名により指定するデータ入出力要求を受け付ける仕組みを提供するファイルシステムが動作していることもあり、この場合、ネットワーク制御部はデータ入出力要求に設定されるファイル名に対応するデータのディスクドライブ上の記憶位置に対応するアドレスを生成し、このアドレスをコマンドに設定する。

【００１５】

本発明の一の態様であるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、前記回路基板に前記ネットワーク制御部および前記ディスク制御部が共通してアクセス可能なメモリを備え、前記ネットワーク制御部および前記ディスク制御部が設定されたタイミングで前記メモリに自身の動作状態を示す動作状態情報を更新する手段を備え、前記動作状態情報に基づいて前記ネットワーク制御部もしくは前記ディスク制御部に障害が発生したことを検知する手段を備えることとしている。

【００１６】

ここで、ネットワーク制御部が、ディスク制御部に送信したコマンドについての受信通知を取得できない場合に動作状態情報に基づいて送信先のディスク制御部の動作状態を調査し、その調査の結果に応じてそのコマンドを再びディスク制御部に送信するかどうかを判断するようにすることもできる。

【００１７】

また、ネットワーク制御部が、ディスク制御部に送信したコマンドについての受信通知を取得できない場合に動作状態情報に基づいて送信先の前記ディスク制御部の動作状態を調査し、ディスク制御装置が正常に動作していないと判断した

場合には、他のディスク制御部に前記コマンドを送信するようにすることもできる。また、障害を検知した場合にその旨を通知するユーザインタフェースを設けてもよい。さらに、障害を検知した場合に、障害が発生している前記ネットワーク制御部もしくは前記ネットワーク制御部に、再起動を要求する信号を送信するように構成してもよい。これによりフェールオーバを実現できる。

【００１８】

このようにネットワーク制御部とディスク制御部とを同一回路基板に形成し、さらに両者が共通してアクセス可能なメモリを設けることで、メモリにネットワーク制御部およびディスク制御部の動作状態情報を記憶しネットワーク制御部およびディスク制御部における障害を検知するようにするといった障害検知の仕組みを構成することができる。

【００１９】

そしてこの構成では、ネットワーク制御部やディスク制御部からの動作状態情報の更新、動作状態情報の参照、再起動信号の送信などの仕組みは、高速かつ信頼性の高い内部バスを用いた通信を介して行われる。このため、ストレージシステムの性能および信頼性を向上させることができる。

【００２０】

また、ネットワーク制御部が、前記コマンドをディスク制御部に送信するに際し前記コマンドの送信先となるディスク制御部の動作状態を動作状態情報から取得し、取得した動作状態に応じて前記コマンドをディスク制御部に送信するかどうかを決定するようにすることもできる。

【００２１】

本発明の一の態様であるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、ディスク制御部はバックアップ装置との接続手段を備え、ネットワーク制御部が、前記ディスクド

ライブに記憶しているデータについてのバックアップ要求を外部装置から受信して前記ディスク制御部にバックアップコマンドを送信し、前記ディスク制御部が、前記バックアップコマンドを受信すると前記ディスクドライブのデータについてのバックアップ指示をバックアップ装置に送信する手段を備えることとしている。

【００２２】

従来、バックアップ装置はファイバチャネルを介してサーバコンピュータに接続する構成が一般的であり、この場合、サーバコンピュータはファイバチャネルを介してディスク制御装置からバックアップ装置にデータ転送を行う必要があった。しかしながら、このようにネットワーク制御部とディスク制御部とが内部バスにより接続する構成では、ネットワーク制御部からバックアップ指示コマンドをディスク制御部に送信するのみでディスク制御部によりバックアップを開始させることができる。そして、この構成ではディスク制御部からバックアップ装置に直接にバックアップ対象となるデータの転送が行われ、バックアップ中にネットワーク制御部にかかる負荷は大幅に低減されることになる。また、ファイバチャネル等の通信を介さない分、処理の高速化が図られる。

【００２３】

本発明の一態様であるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うディスク制御装置において、互いに通信可能に接続された複数の前記回路基板を備え、前記回路基板間でハートビートメッセージを交換することで１の前記回路基板に障害が生じたことを他の回路基板が検知する手段を備え、前記回路基板が他の前記回路基板についての障害を検知した場合に、障害となっている前記回路基板が行っている処理をその回路基板とは異なる他の回路基板に代行させる手段を備えることとしている。

【００２４】

このように、ネットワーク制御部とディスク制御部とが形成された回路基板を複数有する構成では、これら回路基板の間でハートビートメッセージを交換し、他の回路基板に障害が生じた場合にその回路基板で行われていた処理を他の回路基板に代行させるようにすることで、フェールオーバの仕組みを実現することができ、ストレージシステムの信頼性を向上させることができる。

【００２５】

＜システム構成＞

図１に本発明の一実施例として説明するストレージシステムの構成を示す。

ディスク制御装置１０にＬＡＮ５０を介して一台以上のホストコンピュータ２０が接続する。ホストコンピュータ２０は、パーソナルコンピュータやワークステーション、汎用機などである。

【００２６】

ディスク制御装置１０は、ネットワーク制御部１１１、ディスク制御部１１２、障害監視部１１３、キャッシュメモリ１１４、ディスクドライブ１１５などを備えて構成される。なお、ディスクドライブ１１５は、ディスク制御装置１０が収容される筐体とは別筐体に収容されディスク制御装置１０と通信手段を介して接続する構成であってもよい。

【００２７】

このディスク制御装置１０では、ネットワーク制御部１１１、ディスク制御部１１２、障害監視部１１３が、同じ一枚の回路基板１１７上に形成されている。また、この回路基板１１７上には、さらに、ＬＡＮアダプタ（Lan Adaptor）もしくはＮＩＣ（Internet Interface Card）などに相当する機能を実現する回路であるネットワークインタフェース１１８、ＤＭＡコントローラ（Direct Memory Access Controler）１１９が形成されている。

そして、回路基板１１７上に形成されたこれらの回路は、例えば、ＰＣＩバス（Peripheral Component Interconnect Bus）などの内部バス３０により通信可能に結合されている。

【００２８】

ネットワーク制御部１１１は、ＣＰＵ、メモリ（以下、「ネットワーク制御部用メモリ１２１」と称する）などからなる。ネットワーク制御部１１１では、オペレーティングシステムが動作しており、このオペレーティングシステム上では、ＴＣＰ／ＩＰ（登録商標）、ＮＦＳ（Network File System）（登録商標）などのネットワークプロトコルに対応した通信を行うためのプログラムなど、各種のプログラムが動作する。

【００２９】

ディスク制御部１１２は、ＣＰＵ、メモリ（以下、「ディスク制御部用メモリ１２２」と称する）、ディスクドライブ制御回路などからなる。また、ディスク制御部１１２は、ディスクドライブ１１５をＲＡＩＤ（Redundant Array of Inexpensive Disks)の方式で制御する機能を備えていることもある。

障害監視部１１３は、ＣＰＵ、メモリ（以下、「障害監視部用メモリ１２３」と称する）などからなる。

【００３０】

また、ディスク制御装置１０は、前記回路基板１１７上もしくはこれとは別の回路基板上に、バックアップ装置１８４が接続するためのインタフェース回路１８３を備える。なお、このインタフェース回路１８３は内部バス３０に接続されている。インタフェース回路１８３に接続されるバックアップ装置１８４としては、ＤＡＴテープドライブ、ＤＶＤ－ＲＡＭ、ＭＯ、ＣＤ－Ｒ、ディスクドライブ、カセットテープなどがある。

【００３１】

＜データ入出力要求＞

ネットワーク制御部１１１は、ホストコンピュータ２０から送信されてくるデータ入出力要求をＬＡＮ５０を通じて受信すると、図２に示すようにこれをネットワーク制御部用メモリ１２１に確保されている受信バッファに記憶する。なお、図２ではホストコンピュータ２０ごとに別の受信バッファを用意しているが、複数のホストコンピュータ２０に共通の受信バッファを用意してもよい。

【００３２】

ホストコンピュータ２０から送られてくるデータ入出力要求には、それがデー

タ書き込み（Write）要求であるのか、データ読み出し（Read）要求であるのかなどを識別するコマンドコード、ファイル名、処理対象となるデータのファイル内の位置を特定するためのファイル内オフセットアドレスとデータサイズ、そのデータ入出力要求がデータ書き込み要求である場合に設定される書き込みデータ、などの情報が含まれている。

【００３３】

ネットワーク制御部１１１は、受信バッファに記憶しているデータ入出力要求に基づいてＩ／Ｏコマンドを生成し、生成したＩ／Ｏコマンドを、内部バス３０を通じてディスク制御部１１２に送信する。ディスク制御部１１２は、ネットワーク制御部１１１からＩ／Ｏコマンドを受信すると、このＩ／Ｏコマンドに対応するデータ入出力をディスクドライブ１１５に対して行う。

【００３４】

＜ファイル管理情報＞

ネットワーク制御部１１１からディスク制御部１１２に送信されるＩ／Ｏコマンドは、データ入出力要求と、ファイル管理情報に基づいて生成される。ファイル管理情報はディスクドライブ１１５に記憶されている。

【００３５】

図３にファイル管理情報の一例を示す。ファイル管理情報には、ファイル名に対応させて、そのファイルのデータサイズであるファイルサイズ、当該ファイル名に対応するデータが複数の記憶領域に分割されて記憶されている場合における分割数を示すデータ領域数、各連続領域のディスクドライブ１１５上の記憶位置を特定するデータアドレスやデータサイズなどが設定されている。なお、図４に示すようにファイル管理情報は、処理性能の向上のため、ネットワーク制御部用メモリ１２１に記憶（キャッシング）されることもある。

【００３６】

＜Ｉ／Ｏコマンド＞

図５は、データ入出力要求に基づいてＩ／Ｏコマンドが生成される仕組みを説明している。ネットワーク制御部１１１は、まずデータ入出力要求のファイル内オフセットアドレスをファイル管理情報に対照し、そのデータ入出力要求が処理

対象とするファイルに関するデータのディスクドライブ１１５上の記憶位置に対応するデータアドレスとデータサイズとを求める。前述したように、１のファイルに対応するデータは、ディスクドライブ１１５上の連続する記憶領域に纏まって記憶されている場合もあるし、複数の記憶領域に分割されて記憶されている場合もある。図５は、１のファイルに対応するデータが２つの記憶領域に分割されて記憶されている場合である。

【００３７】

ここでデータ入出力要求の処理対象であるファイルが、ディスクドライブ１１５上の連続する記憶領域に纏まって記憶されている場合、ネットワーク制御部１１１はそのデータ入出力要求に対応するＩ／Ｏコマンドとして、その記憶領域を指定するための、１の論理アドレス（ＬＢＡ（Logical Block Address））とブロック数とが設定されたＩ／Ｏコマンドを生成する。ここで論理アドレスとは、ディスクドライブの記憶領域に区画設定された論理的な記憶領域（以下、「論理ユニット」もしくは「ＬＵ（Logical Unit）」と称する）におけるデータの記憶位置を指定する論理的なアドレスである。

【００３８】

一方、データ入出力要求の処理対象であるファイルが、ディスクドライブ１１５の複数の記憶領域に分割されて記憶されている場合、ネットワーク制御部１１１は、分割された各記憶領域のそれぞれを指定する複数の論理アドレスとブロック数との組合せを設定したＩ／Ｏコマンドを生成する。またこの場合には、ファイルがいくつの記憶領域に分割されているかを示す分割数がＩ／Ｏコマンドのリスト数の欄に設定される。

【００３９】

図６は、生成されたＩ／Ｏコマンドがネットワーク制御部用メモリ１２１に記憶されている様子を示している。なお、この図ではＩ／ＯコマンドはＬＵごとにネットワーク制御部用メモリ１２１上に用意されたコマンドテーブルに管理されているが、これに限られるわけではない。

【００４０】

＜Ｉ／Ｏコマンドの生成＞

つぎにネットワーク制御部１１１がデータ入出力要求に基づいてＩ／Ｏコマンドを生成する処理について、図７に示すフローチャートとともに説明する。

ネットワーク制御部１１１は、受信バッファにデータ入出力要求が存在する場合（S711）、受信バッファからデータ入出力要求を一つ取り出し（S712）、データ入出力要求に設定されているファイル名に対応するファイルをネットワーク制御部用メモリ１２１のファイル管理情報から検索する（S713）。

【００４１】

ここで対応するファイルがネットワーク制御部用メモリ１２１のファイル管理情報に存在する場合（S714:YES）には、（S716）に進む。他方、対応するファイルが存在しない場合（S714:NO）はファイル管理情報をディスクドライブ１１５から読み出す（S715）。

【００４２】

（S716）ではデータ入出力要求のオフセットアドレスとデータサイズに対応するディスクドライブ１１５のデータアドレスとデータサイズとを求める。ここで前述したようにそのファイルを構成するデータがディスクドライブ１１５上の複数の記憶領域に分割されて記憶されている場合には、それぞれの領域に対応するデータアドレスとデータサイズとを求める。そして算出したデータアドレスとデータサイズとが設定されたＩ／Ｏコマンドをネットワーク制御部用メモリ１２１上に生成する（S717）。

【００４３】

Ｉ／Ｏコマンドがネットワーク制御部用メモリ１２１に生成される様子を図８に示している。図８（ａ）は、複数のホストコンピュータ２０から１の論理ボリュームについてのＩ／Ｏコマンドが生成される場合におけるＩ／Ｏコマンドの生成の様子である。図８（ｂ）は１のホストコンピュータ２０が複数の論理ボリュームにアクセスする場合におけるＩ／Ｏコマンドの生成の様子である。

【００４４】

図７の（S718）では、ネットワーク制御部１１１は、つぎに処理すべきデータ入出力要求が存在するかどうかを受信バッファから調べる。他のホストコンピュータ２０の受信バッファが存在する場合には、（S712）からの処理が繰り返され

る。

一方、つぎに処理すべきデータ入出力要求が受信バッファに存在しない場合（S578:NO）、ネットワーク制御部１１１は、Ｉ／Ｏコマンドの実行要求を適宜なタイミングでディスク制御部１１２に送信する（S719）。

なお、（S719）の処理は（S718）の処理を待たずに行うように構成することもできる。また以上の処理は、受信バッファが複数存在する場合には各受信バッファについて行われる（S719）。

【００４５】

＜Ｉ／Ｏコマンドの実行＞

つぎに、ディスク制御部１１２によるＩ／Ｏコマンドの実行について、図９に示すフローチャートとともに説明する。

ディスク制御部１１２は、ネットワーク制御部１１１から受信したＩ／Ｏコマンドの実行要求を、ディスク制御部用メモリ１２２に記憶している。ディスク制御部１１２は、ディスク制御部用メモリ１２２にＩ／Ｏコマンドの実行要求が存在するかどうかを監視している（S911）。ディスク制御部１１２は、ディスク制御部用メモリ１２２にＩ／Ｏコマンドの実行要求が存在する場合、ディスク制御部用メモリ１２２からＩ／Ｏコマンドを１つ読み出し（S912）、そのコマンドに設定されているコマンドコードから、そのコマンドがデータ書き込みのコマンドであるのか、データ読み出しのコマンドであるのかを調べる（S913）。ここで読み出しコマンドの場合には、（S914）に進む。一方、データ書き込みコマンドの場合には（S931）に進む。

【００４６】

（S914）においてｎには初期値として１が設定されているものとする。ディスク制御部１１２は、ＬＢＡ番号ｎのＬＢＡおよびデータサイズを読み出す（S915）。つぎにディスク制御部１１２は、そのＬＢＡおよびデータサイズに対応するデータがキャッシュメモリ１１４に存在するかどうかを調べ（S916）、存在する場合（S916:YES）にはキャッシュメモリ１１４からそのデータを読み出してネットワーク用メモリに転送する（S917）。一方、キャッシュメモリ１１４に存在しない場合、ディスク制御部１１２はディスクドライブ１１５に読み出し要求を送

信する（S918）。つぎにディスク制御部１１２はＬＢＡ番号ｎをインクリメントし（S919）、インクリメント後のｎの値をＩ／Ｏコマンドのリスト数と比較する（S920）。ここでｎがリスト数以下の場合（S920:YES）には（S915）に進む。一方、ｎがリスト数よりも大きい場合（S920:NO）には（S921）に進む。

【００４７】

（S921）では、（S918）において読み出し要求を送信した場合（S921:YES）、ディスクドライブ１１５がその要求に対応するデータを読み出すのを待ってから（S923）に進む（S922）。一方、（S918）において読み出し要求を送信していない場合には（S923）に進む。

【００４８】

（S923）では、ディスクドライブ１１５からデータが読み出されると、そのデータをキャッシュメモリ１１４に記憶するとともにネットワーク制御部用メモリ１２１へ転送し、ディスク制御部１１２はネットワーク制御部１１１にデータ入出力要求についての処理を完了した旨のステータスを報告する。

【００４９】

一方、（S913）においてＩ／Ｏコマンドが書き込み要求であった場合には、（S931）に進む。（S931）において、ｎには初期値として１が設定されているものとする。（S932）において、ディスク制御部１１２はＬＢＡ番号ｎのＬＢＡおよびデータサイズを読み出す。

【００５０】

つぎにディスク制御部１１２は、そのＬＢＡおよびデータサイズに対応するデータがキャッシュメモリ１１４に存在するかどうかを調べ（S933）、存在しない場合（S933:NO）には、キャッシュメモリ１１４に記憶領域を確保し（S934）、その記憶領域に書き込みデータを書き込む（S735）。

一方、存在する場合（S933:YES）には、そのデータに書き込みデータを上書きする（S936）。

（S937）では、その書き込みデータに対応させてキャッシュメモリに記憶されている当該書き込みデータについてのデステージフラグをオンにする。なお、デステージフラグがオンになっている書き込みデータは、適宜なタイミングで、キ

ャッシュメモリ１１４からディスクドライブ１１５にデステージされる。

【００５１】

つぎにディスク制御部１１２はＬＢＡ番号ｎをインクリメントし（S938）、インクリメント後のｎがＩ／Ｏコマンドのリスト数に設定されている値と比較する（S939）。ここでｎがリスト数以下の場合（S939:YES）には（S932）に進む。また、ｎがリスト数よりも大きい場合（S939:NO）には（S923）に進み、ディスク制御部はネットワーク制御部にデータ入出力要求についての処理を完了した旨のステータスを報告する（S940）。

【００５２】

以上のように、ネットワーク制御部１１１からディスク制御部１１２に複数のアドレスが設定されている１のコマンドを送信するだけで、ディスク制御部１１２が複数のアドレスに対応するデータ入出力をディスクドライブ１１５に対して行う。このためコマンドの処理にかかるオーバーヘッドが削減されてストレージシステムの性能向上が図られる。

【００５３】

＜バックアップ処理＞

つぎにホストコンピュータ２０から送信されてくるバックアップ要求に応じてディスク制御装置１０が実行するデータバックアップ処理について説明する。

ネットワーク制御部１１１で動作するオペレーティングシステム上では、ホストコンピュータ２０から送信されてくるバックアップ要求を受信した場合にディスクドライブ１１５に記憶しているデータを、図１に示すバックアップ装置１８４のバックアップメディアに記憶する処理を実行する、バックアッププログラムが動作している。

【００５４】

バックアップ要求には、ファイル名や論理ユニット名などの情報により、バックアップ対象となるデータが指定されている。バックアップ要求を受信すると、バックアッププログラムはバックアップ要求に指定されている情報に基づいて、バックアップ対象となるデータのディスクドライブ１１５上の記憶位置を特定するアドレス及びデータサイズを求め、これらを設定したバックアップ指示コマン

ドを内部バス３０を通じてディスク制御部１１２に送信する。

【００５５】

ディスク制御部１１２は、バックアップ指示コマンドを受信すると、このコマンドに指定されているアドレスとデータサイズとで指定される記憶領域に格納されているデータと、バックアップの開始を指示するバックアップ開始コマンドとをバックアップ装置１８４に送信する。バックアップ開始コマンドを受信したバックアップ装置１８４は、このコマンドと共に送られてくるバックアップ対象のデータをバックアップメディアに記録する。

【００５６】

ディスク制御部１１２は、バックアップ対象データのバックアップメディアへの記録が終了した旨の通知をバックアップ装置１８４から受信すると、内部バス３０を通じてバックアッププログラムに完了通知を送信する。バックアッププログラムは完了通知を受信すると、ホストコンピュータ２０にバックアップ完了報告を送信する。

【００５７】

ところで、従来、バックアップ装置１８４はファイバチャネルを介してサーバコンピュータに接続する構成が一般的であり、この場合、サーバコンピュータはファイバチャネルを介してディスク制御装置からバックアップ装置にデータ転送を行う必要があった。

【００５８】

しかしながら、本実施例のようにネットワーク制御部１１１とディスク制御部１１２とが内部バス３０により接続する構成では、ネットワーク制御部１１１からバックアップ指示コマンドをディスク制御部１１２に送信するのみでディスク制御部１１２によりバックアップを開始させることができる。そして、この構成ではディスク制御部１１２からバックアップ装置１８４に直接的にバックアップ対象となるデータの転送が行われ、バックアップ中にネットワーク制御部１１１にかかる負荷は大幅に軽減されることになる。また、ファイバチャネル等の通信を介さない分、処理の高速化も図られる。

【００５９】

＜複数の回路基板＞

フェールオーバさせること等を目的として、ディスク制御装置１０に同一の機能を備える複数の回路基板が実装されることが行われているが、ネットワーク制御部１１１とディスク制御部１１２とが形成された１の回路基板を、１台のディスク制御装置１０に複数枚実装することもできる。図１０は、そのように複数枚の前記回路基板が実装されたディスク制御装置１０を用いて構成したストレージシステムである。

【００６０】

図１０において、各回路基板１１７上には、ネットワークインタフェース１１８、ネットワーク制御部１１１、ディスク制御部１１２およびＤＭＡコントローラ１１９が形成されており、これらはＰＣＩバスなどの内部バス３０に接続している。各回路基板１１７の内部バス３０は互いに接続されている。

【００６１】

また、ディスク制御装置１０には、各回路基板１１７とは別に、障害監視部１１３および共用メモリ１４１が形成された回路基板１２７が実装されている。なお、この回路基板１２７の障害監視部１１３および共用メモリ１４１も、接続ライン３１を介して内部バス３０に接続している。

【００６２】

ここでこのような構成からなるディスク制御装置１０では、各回路基板１１７に形成されているネットワーク制御部１１１間で、内部バス３０を通じてハートビートメッセージを交換して互いに動作状態を監視することで、フェールオーバの仕組みを実現することができる。

【００６３】

以下、この仕組みについて、図１１～図１３に示すフローチャートとともに説明する。

まず、前提として、共用メモリ１４１には、各回路基板のネットワーク制御部１１１の動作状態に関するステータス情報が記憶されているものとする。なお、動作に関するステータス情報とは、例えば、ネットワーク制御部１１１が、正常に動作しているかどうかを示すステータス情報である。

【００６４】

このステータス情報は、定期的もしくは不定期などの設定されたタイミングで、各回路基板１１７のネットワーク制御部１１１により内部バス３０を通じて書き込まれる。なお、このステータス情報は、障害監視部１１３が内部バス３０を通じて各ネットワーク制御部１１１に動作状態を問い合わせ、その応答として取得した各ネットワーク制御部１１１の動作状態を、障害監視部１１３が内部バス３０を通じて間接的に共用メモリ１４１に書き込むこともある。

【００６５】

図１１は、ある回路基板１１７のネットワーク制御部１１１が、他の回路基板１１７の他のネットワーク制御部１１１に対し、ハートビートメッセージを要求する側として動作する場合における当該ネットワーク制御部１１１の処理を説明するフローチャートである。

【００６６】

ネットワーク制御部１１１は、内部バス３０を通じて共用メモリ１４１にアクセスしてステータス情報を調べ、正常に動作しているネットワーク制御部１１１を探索する（S1111）。そして、正常に動作している他のネットワーク制御部１１１が探索されると、そのネットワーク制御部１１１に対し、内部バス３０を通じてハートビートメッセージを要求するメッセージを送信し（S1112）、そのネットワーク制御部１１１からハートビートメッセージが送られてくるのを一定時間待つ（S1113）。

【００６７】

ここで他のネットワーク制御部１１１からハートビートメッセージが送られてきた場合（S1114：YES）には、リトライカウンタに「０」を設定し（S1115）、一定時間待機した後（S1116）、再び（S1111）からの処理に戻る。

【００６８】

一方、（S1114）において、他のネットワーク制御部１１１からハートビートメッセージが受信できなかった場合（S1114：NO）には、リトライカウンタに「１」を加算し（S1117）、リトライカウンタをあらかじめ設定されているリトライオーバ閾値と比較する（S1118）。

【００６９】

ここでリトライカウンタの値がリトライオーバ閾値を超えている場合（S1118：YES）には、内部バス３０を通じて障害監視部１１３に、他のネットワーク制御部１１１に障害が発生している旨のメッセージを送信する（S1119）。

なお、このメッセージには、その障害がハードウエアの障害に起因するのか、ソフトウエアに起因するのかを示す情報や、あるネットワーク制御部１１１からのハートビートメッセージが中断している旨を通知する情報など、障害の内容に関する情報も付帯される。ネットワーク制御部１１１は、そのメッセージに対して障害監視部１１３から応答があると（S1120）、他のネットワーク制御部１１１の処理を代行する、フェールオーバ処理を開始する(S1121)。

【００７０】

一方、（S1118）の処理において、リトライカウンタの値がリトライオーバ閾値を超えていない場合（S1118：NO）には、リトライカウンタに「０」を設定し（S1115）、再び（S1111）からの処理に戻る。

【００７１】

図１２は、ある回路基板１１７のネットワーク制御部１１１が、ハートビートメッセージを受信したり、障害監視部１１３からのメッセージを受信した場合における処理を説明するフローチャートである。

ネットワーク制御部１１１は、内部バス３０を通じて送られてくるメッセージを受信すると（S1211）、そのメッセージがハートビートメッセージであるかどうかを調べる（S1212）。

ここで受信したメッセージがハートビートメッセージである場合（S1212：YES）、ネットワーク制御部１１１は、そのハートビートメッセージを送信したネットワーク制御部１１１に対し、ハートビートメッセージを送信する（S1213）。

【００７２】

一方、受信したメッセージがハートビートメッセージ以外のメッセージであった場合（S1212：YES）は、ネットワーク制御部１１１はそのメッセージが障害監視部１１３から送られてくる、他のネットワーク制御部１１１が動作していない旨を通知するメッセージであるかどうかを調べる（S1214）。

（S1214）において、そのメッセージが、他のネットワーク制御部１１１が動作していない旨を通知する障害監視部１１３からのメッセージであった場合には（S1214：YES）、そのネットワーク制御部１１１との間の通信を抑止して（S1215）、必要な場合にはそのネットワーク制御部１１１に関するフェールオーバ処理を開始する（S1216）。

【００７３】

図１３は、障害制御部１３の処理を説明するフローチャートである。障害制御部１３は、図１１における（S1119）の通知、すなわち、あるネットワーク制御部１１１に障害が発生している旨のメッセージを受信すると（S1311：YES）、そのメッセージに付帯する情報に基づいて、その障害通知がハードウエアの障害によるものであるのかを判断する（S1312）。

【００７４】

（S1312）において、ハードウエアの障害であると判断した場合には（S1312：YES）、障害監視部１１３は、前記メッセージにより特定されるネットワーク制御部１１１が動作していない旨のステータス情報を共用メモリ１４１に書き込む（S1313）。そして、障害監視部１１３は、前記メッセージで特定されるネットワーク制御部１１１に障害が発生している旨を記載したメッセージを、障害が発生してないネットワーク制御部１１１に送信する（S1314）。

【００７５】

一方、（S1312）において、ハードウエアの障害でないと判断した場合には（S1312：NO）、障害監視部１１３は、受信したメッセージが、ネットワーク制御部１１１からハートビートメッセージが送られてきていない旨を通知するメッセージであるかどうかを調べる（S1315）。そして、障害監視部１１３は、受信したメッセージが、ネットワーク制御部１１１からハートビートメッセージが送られてきていない旨を通知するメッセージであった場合には、内部バス３０を通じてそのネットワーク制御部１１１を制御してそのネットワーク制御部１１１の動作を停止させた（S1316）後、（S1313）からの処理に進む。

【００７６】

以上のようにしてフェールオーバの仕組みを実現することができる。ちなみに

、ＬＡＮを介してハートビートメッセージを通信する場合には、ハートビートメッセージの通信によりＬＡＮに負荷がかかり、また、この負荷を回避しようとすれば、ハートビートメッセージのための専用のＬＡＮを設ける必要があるが、内部バス３０によりハートビートメッセージを通信する仕組みとした場合にはこのような問題や煩わしさが一切生じないという利点がある。

【００７７】

＜障害監視＞

処理能力の向上や可用性の向上などを目的として、ディスク制御装置１０に複数のネットワーク制御部１１１やディスク制御部１１２が実装されることがある。

このような場合、複数のネットワーク制御部１１１やディスク制御部１１２を、図１４に示すように同一の回路基板１４０上に形成することで、処理能力の向上や信頼性の向上を図ることができる。なお、回路基板１４０には、ネットワーク制御部１１１およびディスク制御部１１２が共通してアクセス可能な共用メモリ１４１が形成されている。

【００７８】

共用メモリ１４１には、ネットワーク制御部１１１やディスク制御部１１２が内部バス３０を通じて書き込む時刻情報を記憶するための記憶領域（以下、「タイマテーブル」と称する）が確保される。タイマテーブルには、ネットワーク制御部１１１やディスク制御部１１２の動作状態に関する情報が設定される。ここでは動作状態情報としてその書き込みが行われた時刻が書き込まれる。また、ネットワーク制御部１１１やディスク制御部１１２からタイマテーブルへの書き込みは、障害監視に必要とされる設定されたタイミング（例えば１秒間隔）で行われる。

【００７９】

図１５は、ネットワーク制御部１１１がディスク制御部１１２にＩ／Ｏコマンドを送信するに際し、ネットワーク制御部１１１により行われる障害検知の仕組みを説明するフローチャートである。

ネットワーク制御部１１１は、ディスク制御部１１２にＩ／Ｏコマンドを送信

するにあたり、事前にタイマテーブルを参照してそのＩ／Ｏコマンドの送信先のディスク制御部１１２が正常に動作しているかどうかを調査する（S1511,S1512）。なお、この調査は、例えば、タイマの更新が直前の所定秒数以内に行われていたかどうかを調べることにより行われる。

【００８０】

この調査によりＩ／Ｏコマンドを送信しようとするディスク制御部１１２に対応するタイマテーブルの更新が行われていない（正常に動作していない）場合には（S1512:NO）、ネットワーク制御部１１１はその処理を中断してホストコンピュータ２０にエラーを報告し、また正常に動作しているディスク制御部１１２を検索し（S1513）、検索された他のディスク制御部１１２にＩ／Ｏコマンドの送信を試みる（S1514）。

一方、（S1512）においてＩ／Ｏコマンドを送信しようとするディスク制御部１１２についてタイマテーブルの更新が行われている（正常に動作している）場合には（S1512:YES）、そのディスク制御部１１２にコマンドを送信する（S1515）。

【００８１】

ここで前記調査によりネットワーク制御部１１１がＩ／Ｏコマンドの送信先のディスク制御部１１２が正常に動作していると判断（S1512:YES）し、Ｉ／Ｏコマンドをそのディスク制御部１１２に送信した（S1514, S1515）にもかかわらず、そのＩ／Ｏコマンドが送信先に受け付けられず、タイムアウトとなってしまうことがあるが、Ｉ／Ｏコマンドを送信した後にディスク制御部１１２に障害が発生した可能性もあるため、この場合は（S1511）に進む。

一方、（S1516）において、タイムアウトにならずにネットワーク制御部１１１がディスク制御部１１２からＩ／Ｏコマンドの受領通知を受信すれば処理が終了する（S1517）。

【００８２】

以上の障害監視の仕組みによれば、ネットワーク制御部１１１からディスク制御部１１２に対するＩ／Ｏコマンドを確実にディスク制御部１１２に引き渡すことが可能となり、これによりストレージシステムの信頼性を向上させることがで

きる。

【００８３】

＜障害監視部＞

ところで、前出の図１４に示す回路基板１４０には、各ネットワーク制御部１１１および各ディスク制御部１１２に対応するタイマテーブルの時刻が更新されているかどうかを監視する障害監視部１３が形成されている。障害監視部１３は、例えば、一定時間以上、時刻の更新を行っていないネットワーク制御部１１１もしくはディスク制御部１１２を検知し、ＳＮＭＰに従って外部装置やユーザインタフェースに障害の発生を通知する。

【００８４】

また障害監視部１３は、時刻が更新されていない原因がネットワーク制御部１１１やディスク制御部１１２のファームウエアの障害によるものであると判断した場合、ネットワーク制御部１１１やディスク制御部１１に内部バス３０を通じてリセット信号（例えば、Power on reset 信号）を送信する。また、障害監視部１３は、内部バス３０を通じて送られてくるネットワーク制御部１１１やディスク制御部１１２からの通知によっても、ネットワーク制御部１１１やディスク制御部１１２に障害の発生を認知する。

【００８５】

以上のように、ネットワーク制御部１１１、ディスク制御部１１２、共用メモリ１４１、障害監視部１３が、それぞれ同じ回路基板上に形成されていることで、ネットワーク制御部１１１やディスク制御部１１２からの共用メモリ１４１のタイマテーブルへの時刻の書き込み、ネットワーク制御部１１１によるタイマテーブルの参照、障害監視部１３によるタイマテーブルの時刻の更新状態の監視、ネットワーク制御部１１１やディスク制御部１１２へのリセット信号の送信などの処理は、全て内部バス３０を通じて行われることになる。従って、これらの処理は高速かつ高信頼に行われ、ディスク制御装置１０の処理の高速化や信頼性の向上が図られる。

【００８６】

以上、本発明に係るディスク制御装置等について説明してきたが、上記した発

明の実施の形態は、本発明の理解を容易にするためのものであり、本発明を限定するものではない。本発明は、その趣旨を逸脱することなく、変更、改良され得ると共に、本発明にはその等価物が含まれることは勿論である。

【００８７】

【発明の効果】

本発明によれば、ディスク制御装置およびディスク制御装置の制御方法を提供することができる。

【図面の簡単な説明】

【図１】　本発明の一実施例による、ストレージシステムの構成を示す図である。

【図２】　本発明の一実施例による、データ入出力要求がネットワークデータ用メモリの受信バッファに記憶される様子を説明する図である。

【図３】　本発明の一実施例による、ファイル管理情報を示す図である。

【図４】　本発明の一実施例による、ファイル管理情報が制御用メモリに記憶される様子を説明する図である。

【図５】　本発明の一実施例による、データ入出力要求に基づいてＩ／Ｏコマンドが生成される仕組みを説明する図である。

【図６】　本発明の一実施例による、生成されたＩ／Ｏコマンドが制御用メモリに記憶されている様子を説明する図である。

【図７】　本発明の一実施例による、ネットワーク制御部がデータ入出力要求に基づいてＩ／Ｏコマンドを生成する処理を説明するフローチャートである。

【図８】　（ａ）、（ｂ）は、それぞれ本発明の一実施例による、Ｉ／Ｏコマンドが制御用メモリに生成される様子を説明する図である。

【図９】　本発明の一実施例による、Ｉ／Ｏコマンドの実行を説明するフローチャートを示す図である。

【図１０】　本発明の一実施例による、複数枚の回路基板が実装されたディスク制御装置を用いて構成されるストレージシステムの構成を示す図である。

【図１１】　本発明の一実施例による、ある回路基板のネットワーク制御部が、他の回路基板の他のネットワーク制御部に対し、ハートビートメッセージを

要求する側として動作する場合における当該ネットワーク制御部の処理を説明するフローチャートを示す図である。

【図１２】　本発明の一実施例による、ある回路基板のネットワーク制御部が、ハートビートメッセージを受信したり、障害監視部からのメッセージを受信した場合における処理を説明するフローチャートを示す図である。

【図１３】　本発明の一実施例による、障害制御部の処理を説明するフローチャートである。

【図１４】　本発明の一実施例による、複数のネットワーク制御部およびディスク制御部が形成された回路基板を示す図である。

【図１５】　本発明の一実施例による、障害検知の仕組みを説明するフローチャートを示す図である。

【図１６】　従来の典型的なストレージシステムの構成を示す図である。

【符号の説明】

１０　ディスク制御装置

１８　回路基板

２０　ホストコンピュータ

３０　内部バス

５０　ＬＡＮ（ネットワーク）

１１１　ネットワーク制御部

１１２　ディスク制御部

１１３　障害監視部

１１４　キャッシュメモリ

１１５　ディスクドライブ

１２１　ネットワーク制御部用メモリ

１２２　ディスク制御部用メモリ

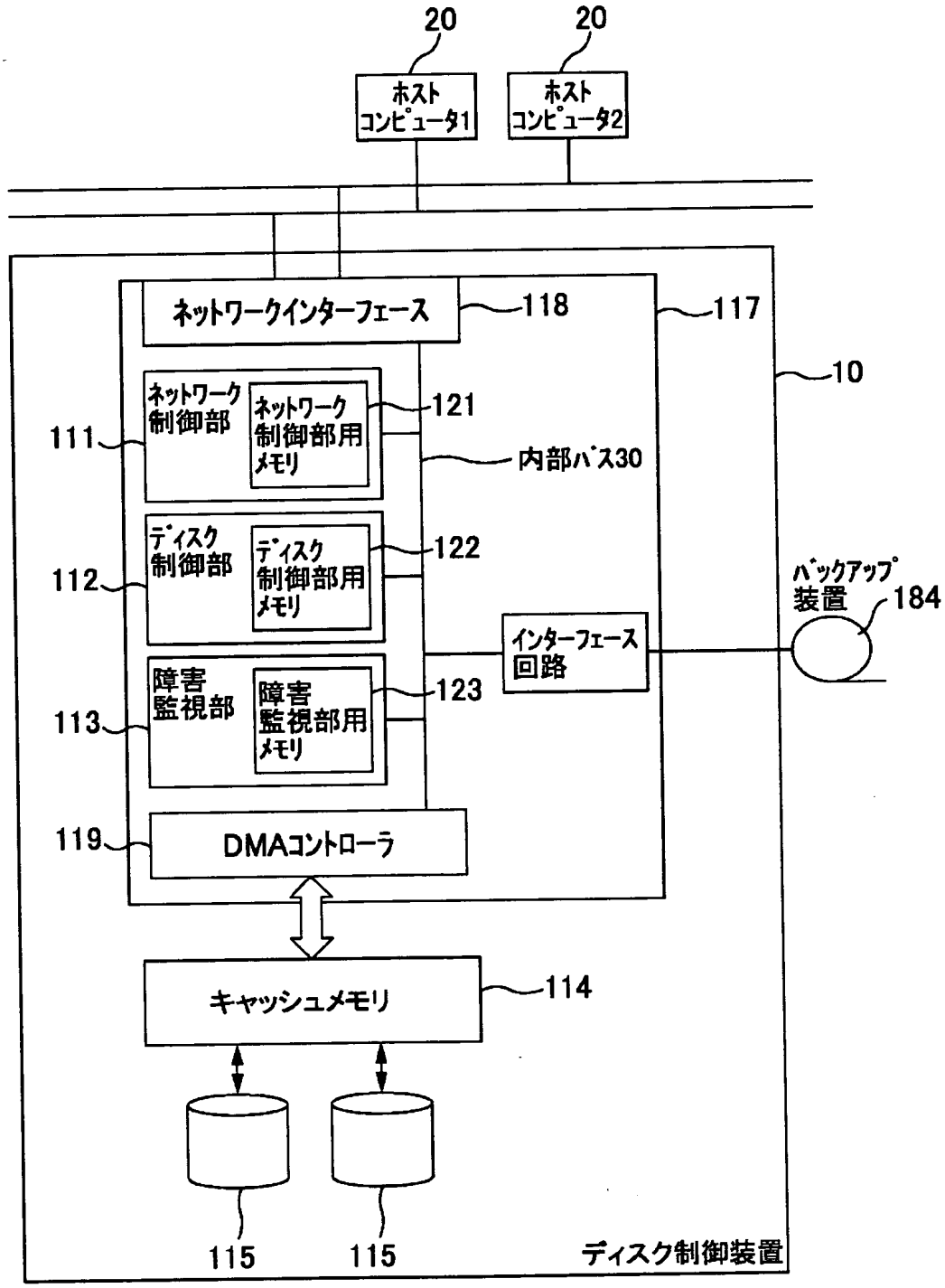
１４０　回路基板

１４１　共用メモリ

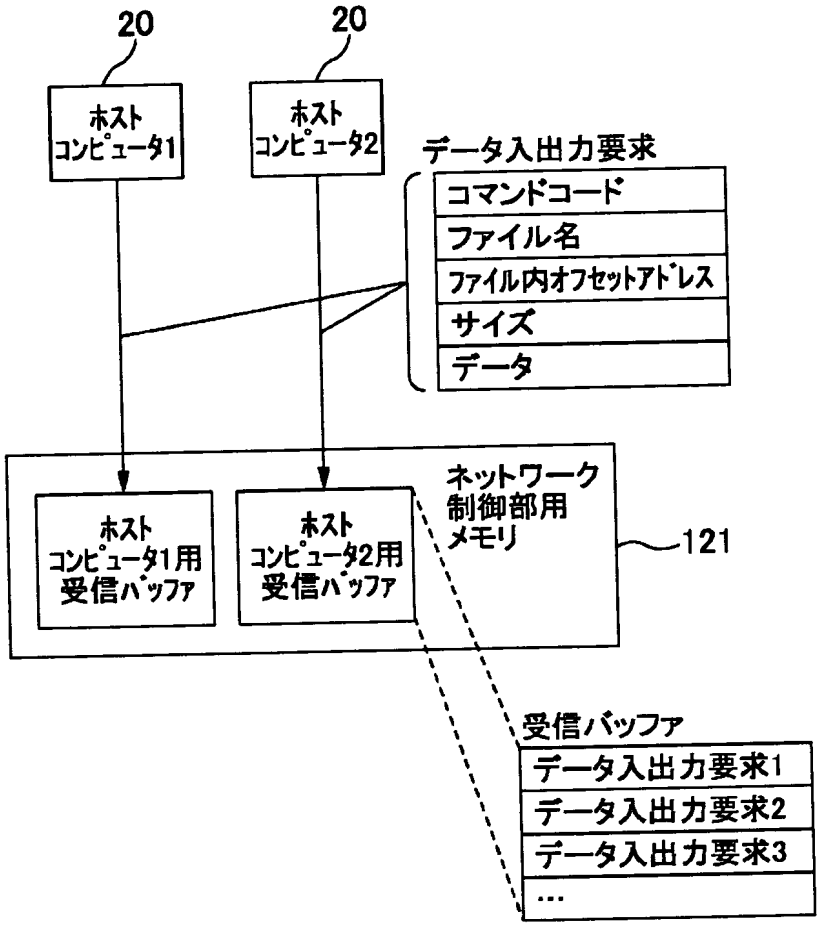
１８４　バックアップ装置

【書類名】　図面

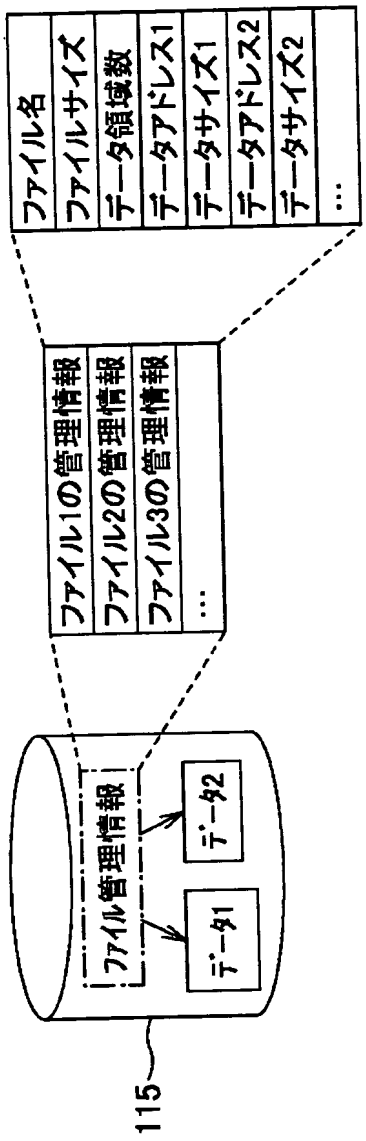
【図１】


20
20
ホスト
コンピュータ1
ホスト
コンピュータ2
ネットワークインターフェース
118
117
10
ネットワーク
制御部
ネットワーク
制御部用
メモリ
121
111
内部バス30
ディスク
制御部
ディスク
制御部用
メモリ
122
112
バックアップ
装置
184
インターフェース
回路
障害
監視部
障害
監視部用
メモリ
123
113
119
DMAコントローラ
キャッシュメモリ
114
115
115
ディスク制御装置

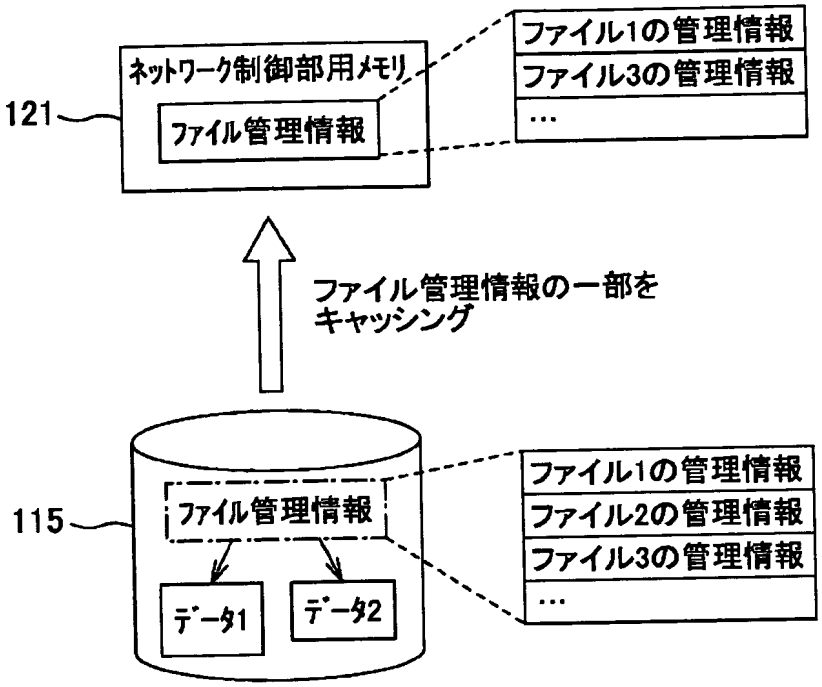
【図２】


20
20
ホスト
コンピュータ1
ホスト
コンピュータ2
データ入出力要求
コマンドコード
ファイル名
ファイル内オフセットアドレス
サイズ
データ
ホスト
コンピュータ1用
受信バッファ
ホスト
コンピュータ2用
受信バッファ
ネットワーク
制御部用
メモリ
121
受信バッファ
データ入出力要求1
データ入出力要求2
データ入出力要求3
…

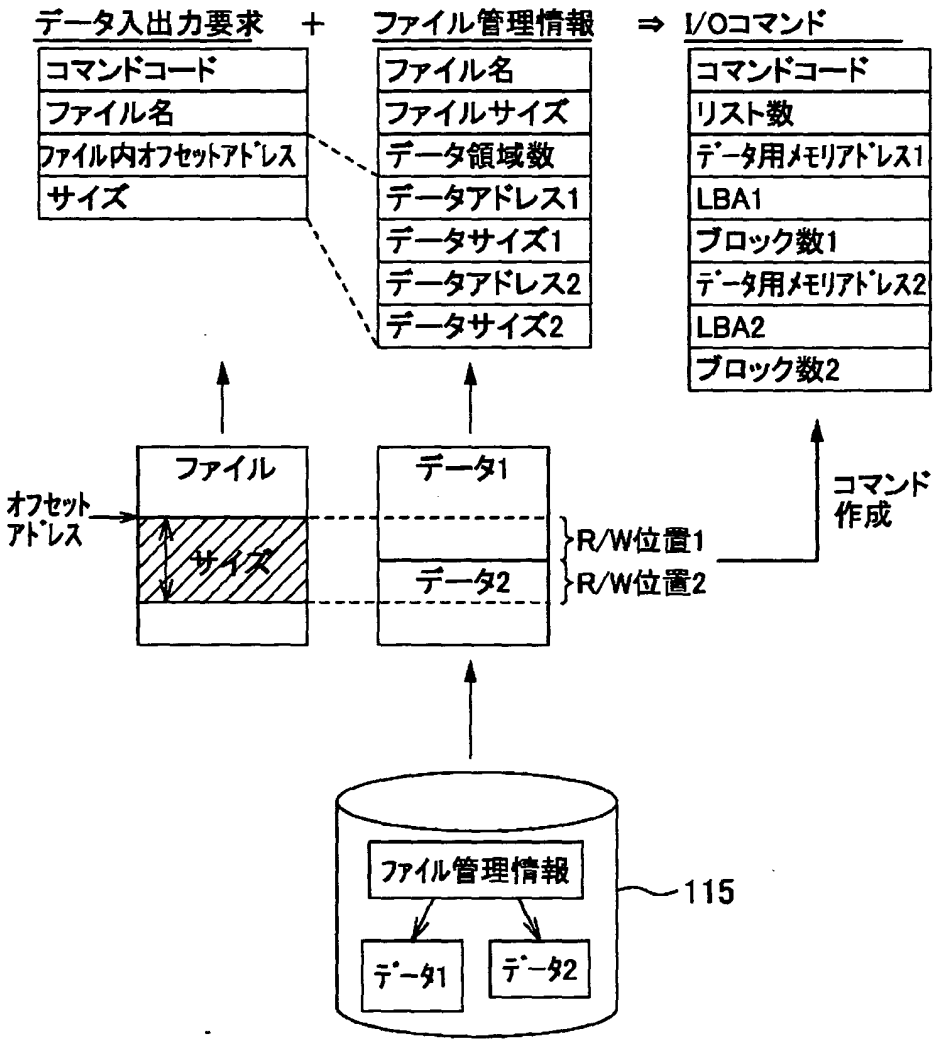
【図３】


ファイル名
ファイルサイズ
データ領域数
データアドレス1
データサイズ1
データアドレス2
データサイズ2
…
ファイル1の管理情報
ファイル2の管理情報
ファイル3の管理情報
…
ファイル管理情報
データ2
データ1
115

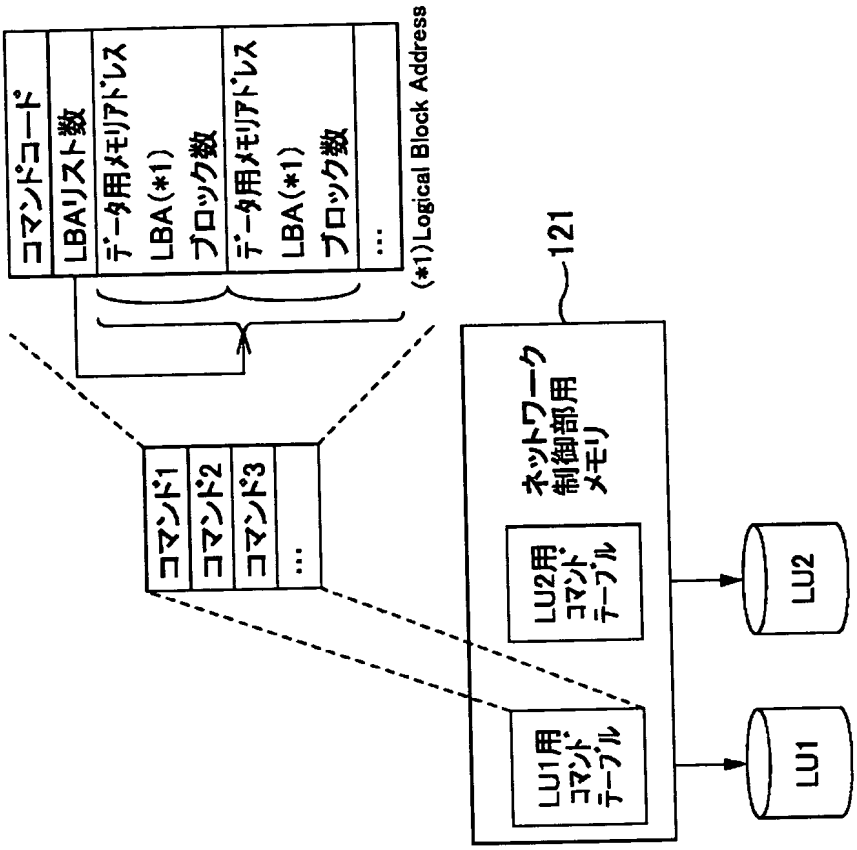
【図４】


ネットワーク制御部用メモリ
ファイル管理情報
121
ファイル1の管理情報
ファイル3の管理情報
…
ファイル管理情報の一部を
キャッシング
ファイル管理情報
115
データ1
データ2
ファイル1の管理情報
ファイル2の管理情報
ファイル3の管理情報
…

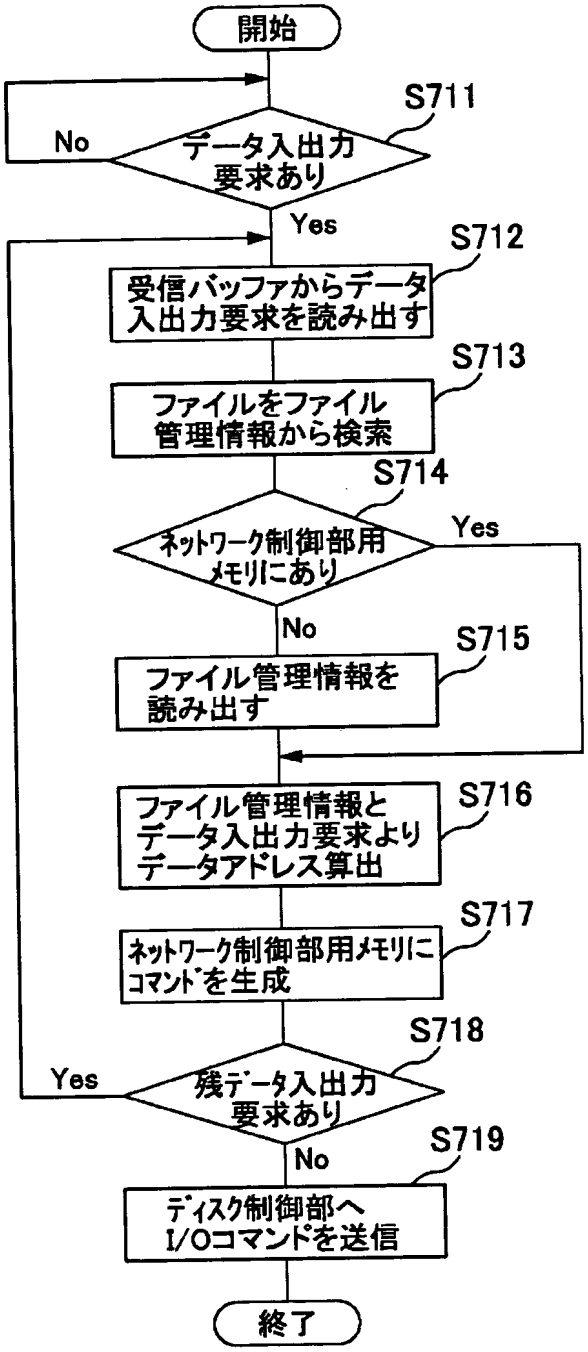
【図５】


データ入出力要求
＋
ファイル管理情報
⇒
I/Oコマンド
コマンドコード
ファイル名
ファイル内オフセットアドレス
サイズ
ファイル名
ファイルサイズ
データ領域数
データアドレス1
データサイズ1
データアドレス2
データサイズ2
コマンドコード
リスト数
データ用メモリアドレス1
LBA1
ブロック数1
データ用メモリアドレス2
LBA2
ブロック数2
ファイル
データ1
オフセット
アドレス
サイズ
R/W位置1
データ2
R/W位置2
コマンド
作成
ファイル管理情報
115
データ1
データ2

【図６】

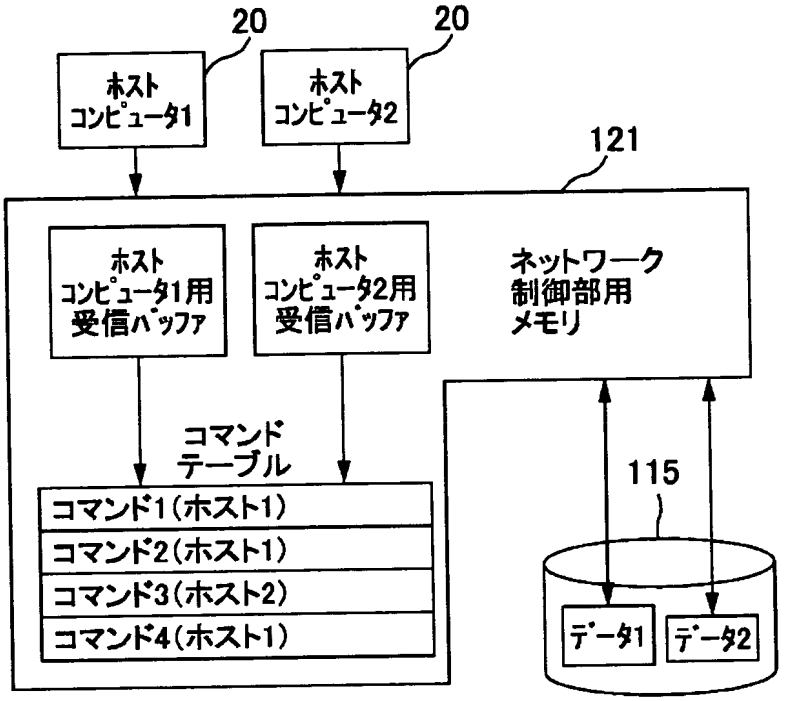

コマンドコード
LBAリスト数
データ用メモリアドレス
LBA(*1)
ブロック数
データ用メモリアドレス
LBA(*1)
ブロック数
...
(*1)Logical Block Address
コマンド1
コマンド2
コマンド3
...
121
ネットワーク制御部用メモリ
LU2用コマンドテーブル
LU1用コマンドテーブル
LU2
LU1

【図７】

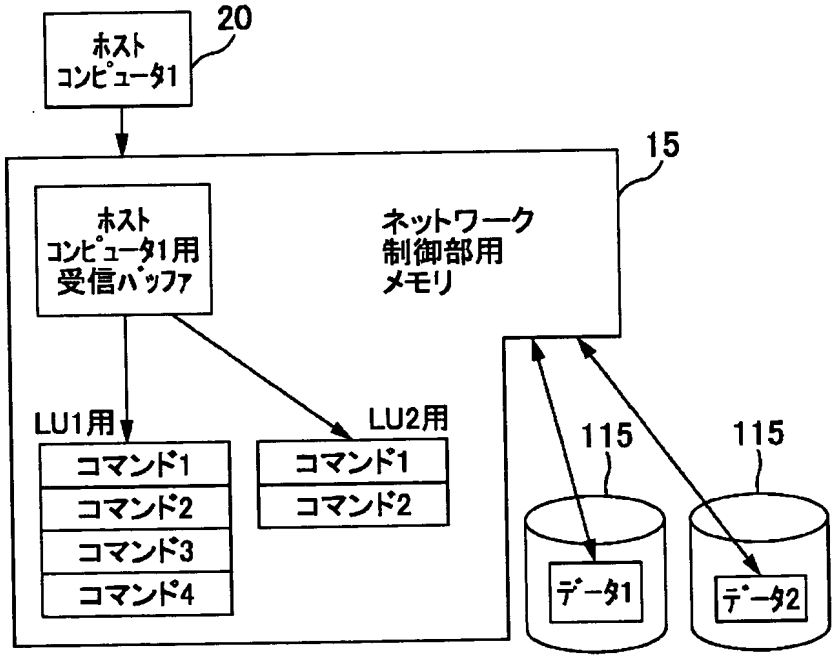

開始
S711
No
データ入出力
要求あり
Yes
S712
受信バッファからデータ
入出力要求を読み出す
S713
ファイルをファイル
管理情報から検索
S714
ネットワーク制御部用
メモリにあり
Yes
No
S715
ファイル管理情報を
読み出す
S716
ファイル管理情報と
データ入出力要求より
データアドレス算出
S717
ネットワーク制御部用メモリに
コマンドを生成
S718
Yes
残データ入出力
要求あり
No
S719
ディスク制御部へ
I/Oコマンドを送信
終了

【図８】

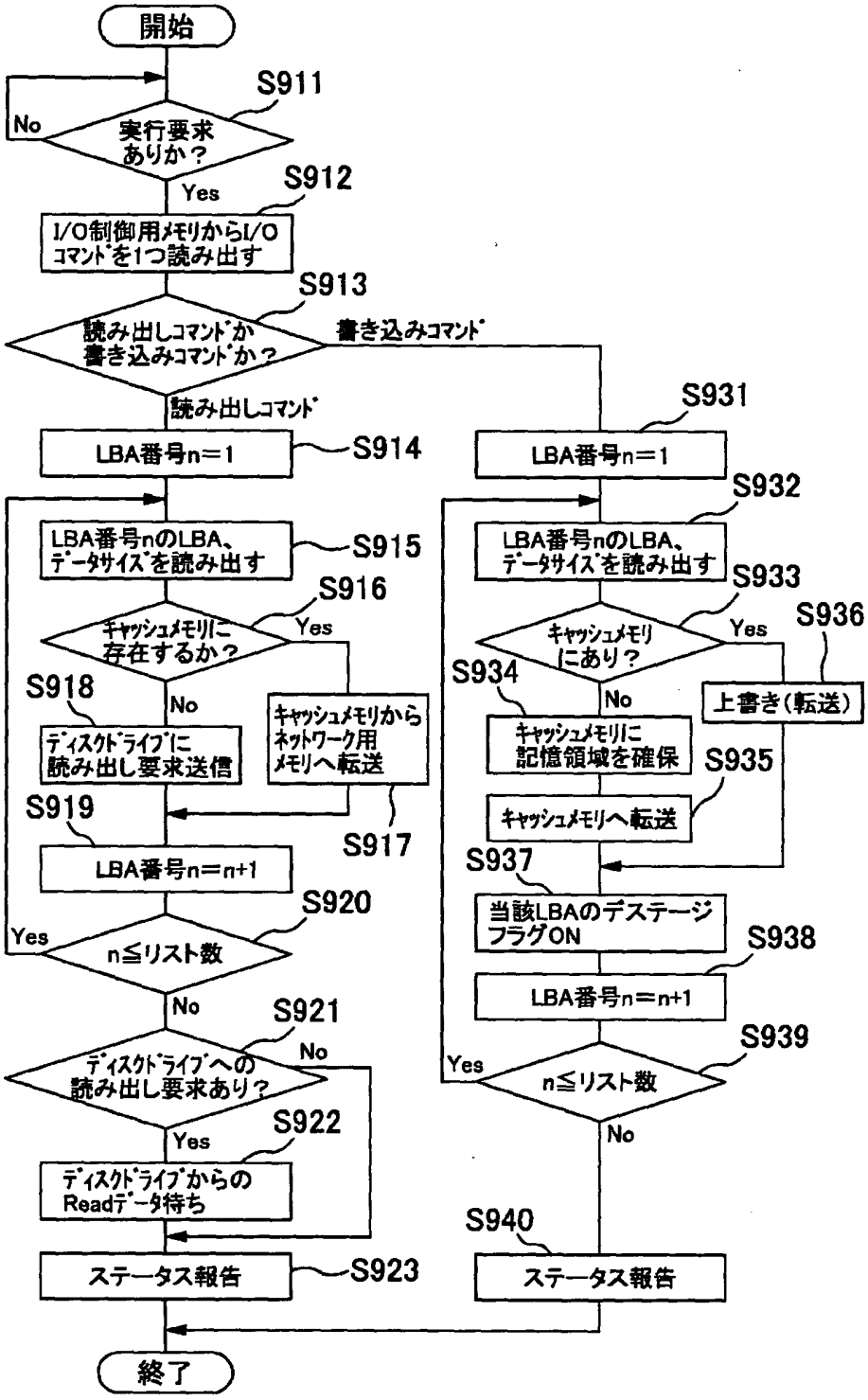
(a)


20
20
ホスト
コンピュータ1
ホスト
コンピュータ2
121
ホスト
コンピュータ1用
受信バッファ
ホスト
コンピュータ2用
受信バッファ
ネットワーク
制御部用
メモリ
コマンド
テーブル
コマンド1(ホスト1)
コマンド2(ホスト1)
コマンド3(ホスト2)
コマンド4(ホスト1)
115
データ1
データ2

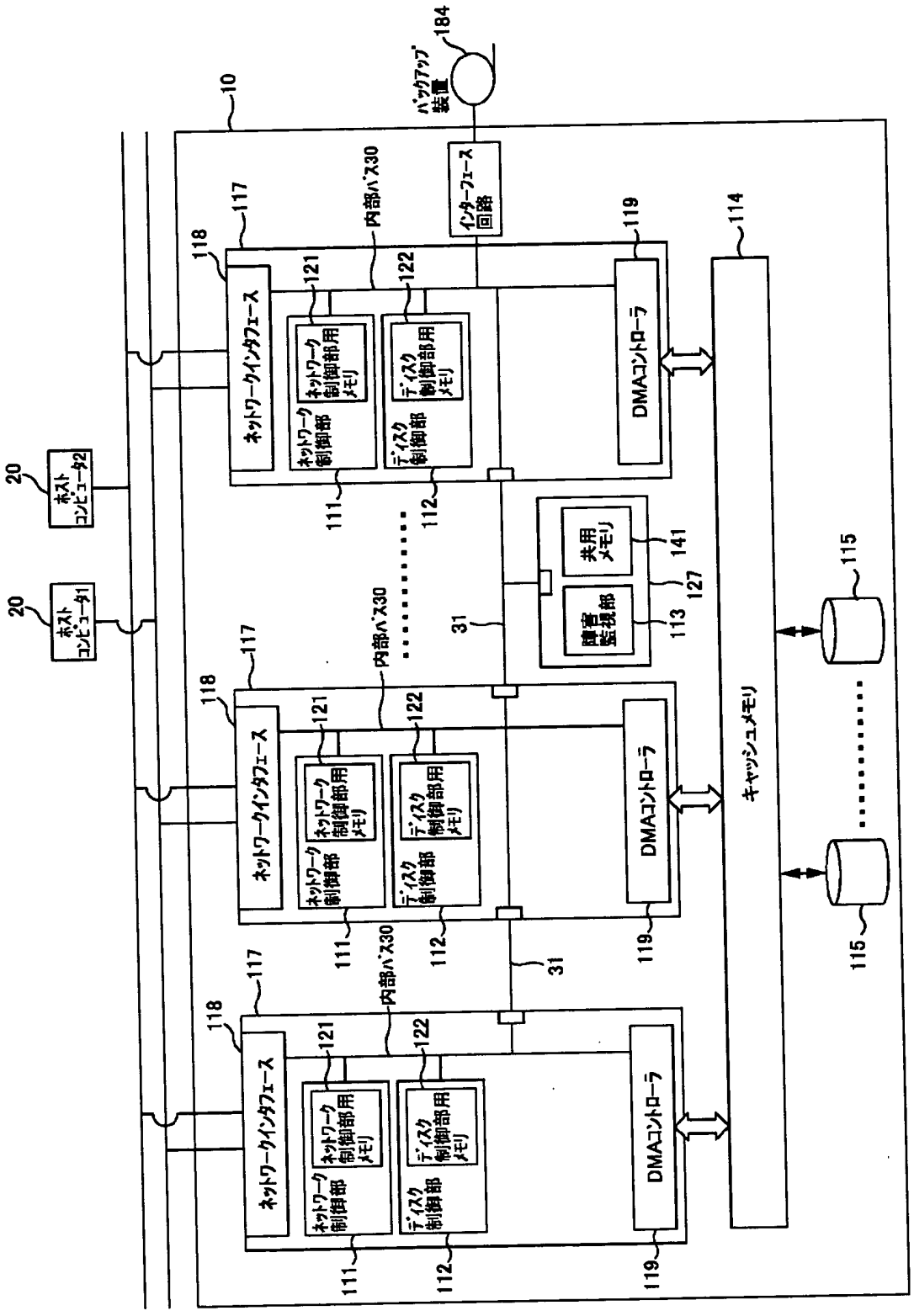
(b)


20
ホスト
コンピュータ1
15
ホスト
コンピュータ1用
受信バッファ
ネットワーク
制御部用
メモリ
LU1用
LU2用
コマンド1
コマンド2
コマンド3
コマンド4
コマンド1
コマンド2
115
115
データ1
データ2

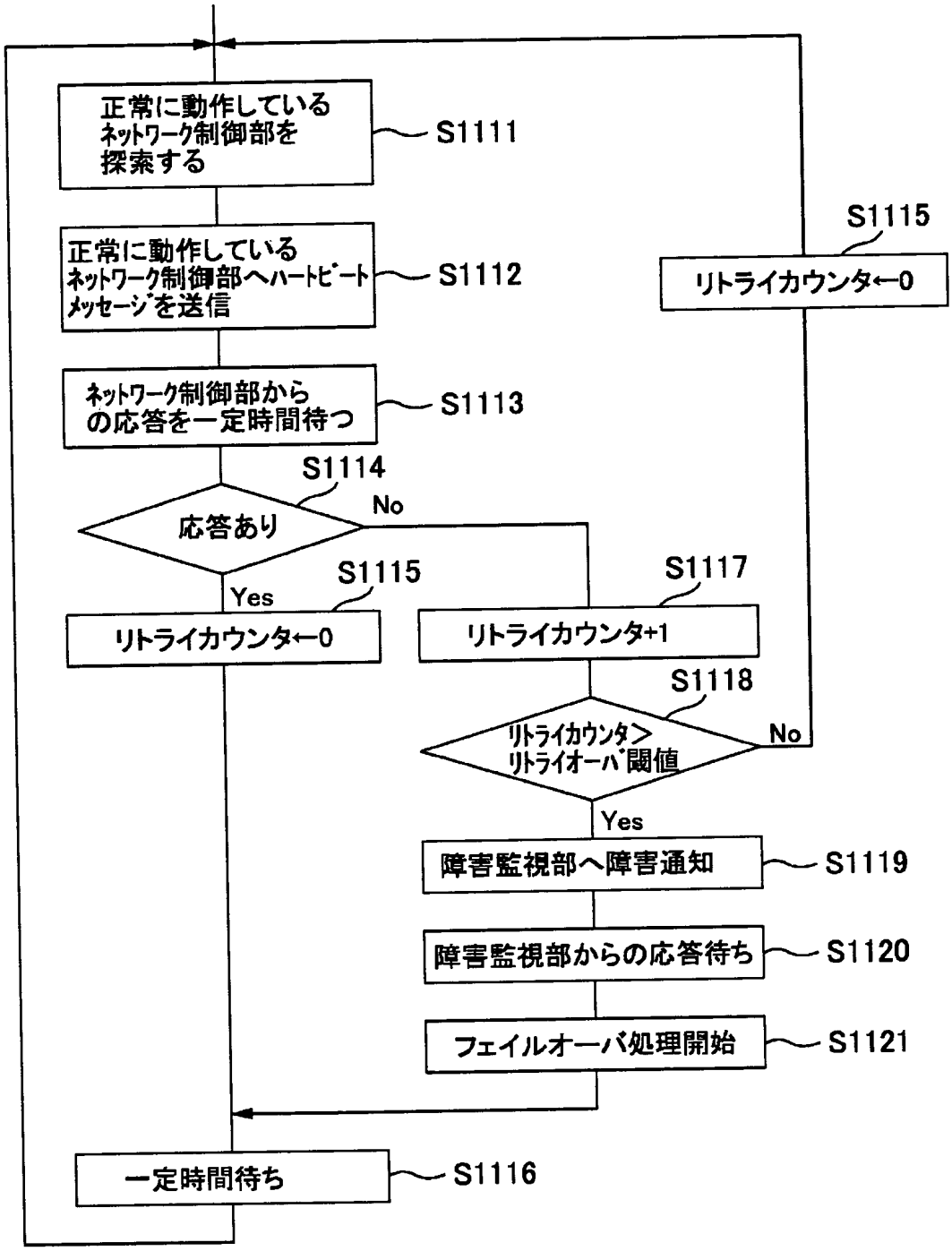
【図９】


開始
S911
No
実行要求ありか？
Yes
S912
I/O制御用メモリからI/Oコマンドを1つ読み出す
S913
読み出しコマンドか書き込みコマンドか？
書き込みコマンド
読み出しコマンド
S914
LBA番号n＝1
S915
LBA番号nのLBA、データサイズを読み出す
S916
キャッシュメモリに存在するか？
Yes
No
S918
ディスクドライブに読み出し要求送信
キャッシュメモリからネットワーク用メモリへ転送
S917
S919
LBA番号n＝n+1
S920
Yes
n≦リスト数
No
S921
No
ディスクドライブへの読み出し要求あり？
Yes
S922
ディスクドライブからのReadデータ待ち
ステータス報告
S923
終了
S931
LBA番号n＝1
S932
LBA番号nのLBA、データサイズを読み出す
S933
キャッシュメモリにあり？
Yes
S936
上書き（転送）
S934
No
キャッシュメモリに記憶領域を確保
S935
キャッシュメモリへ転送
S937
当該LBAのデステージフラグON
S938
LBA番号n＝n+1
S939
Yes
n≦リスト数
No
S940
ステータス報告

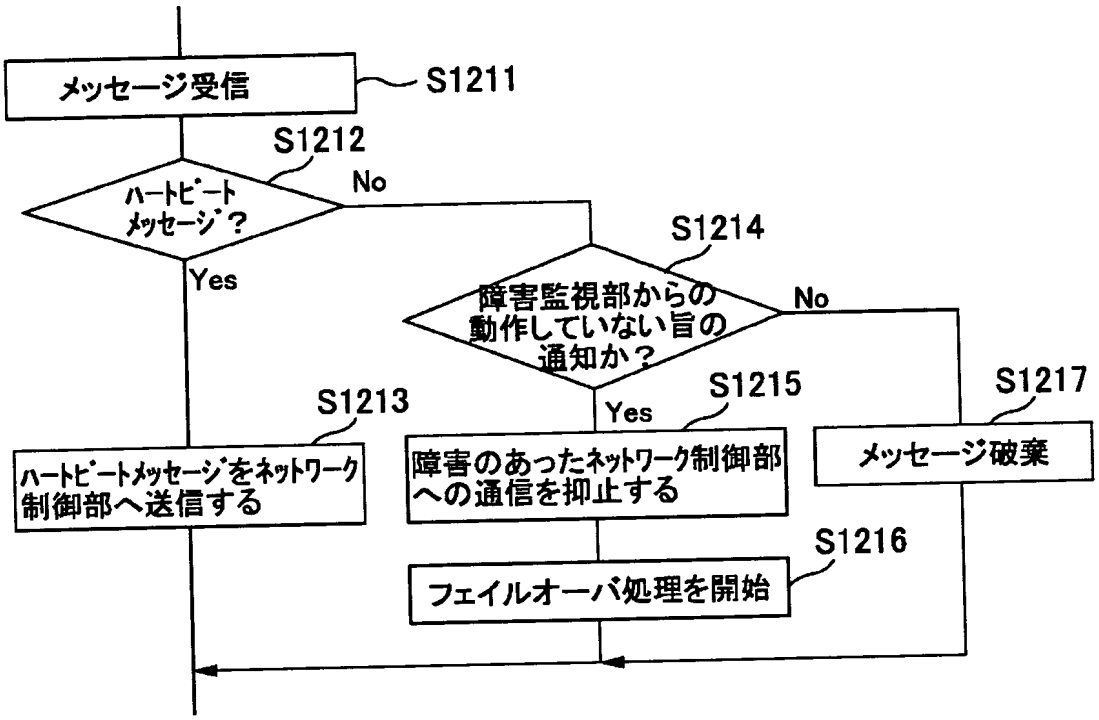
【図１０】


10
184
バックアップ装置
インターフェース回路
内部バス30
117
118
119
121
122
ネットワークインタフェース
ネットワーク制御部
ネットワーク制御部用メモリ
ディスク制御部
ディスク制御部用メモリ
DMAコントローラ
111
112
31
114
キャッシュメモリ
115
127
113
障害監視部
141
共用メモリ
20
ホストコンピュータ1
ホストコンピュータ2

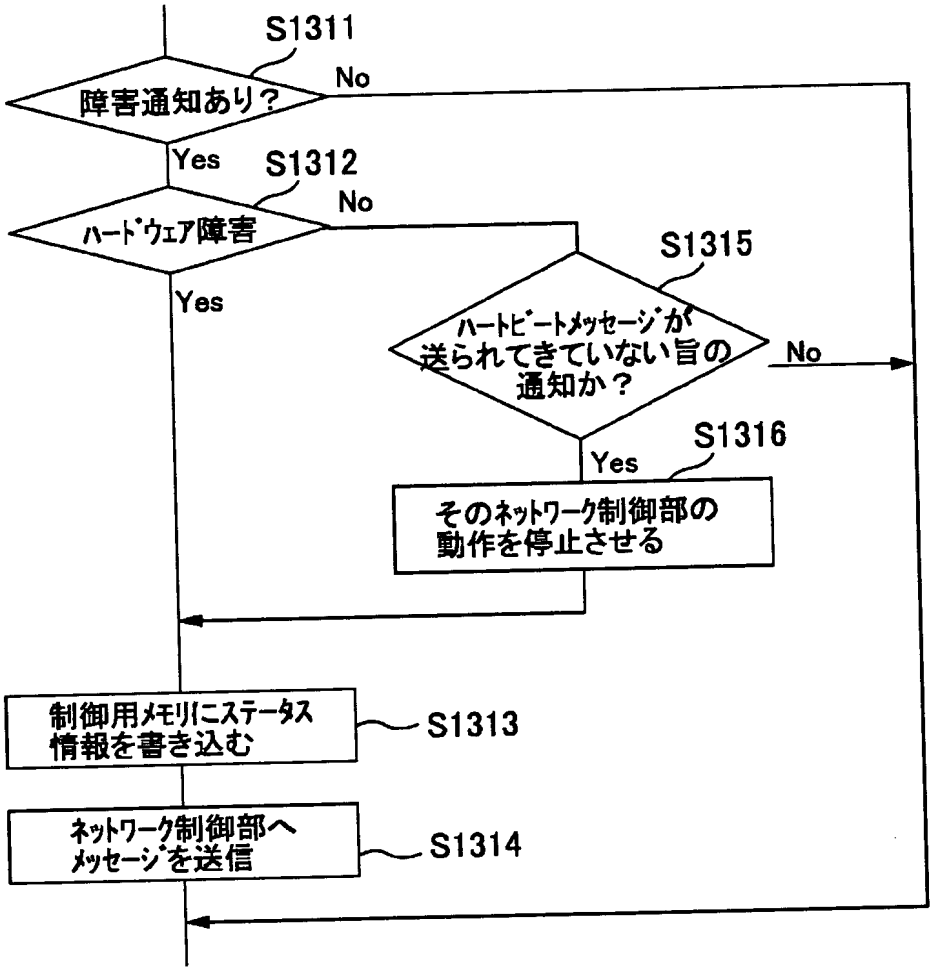
【図１１】


正常に動作している
ネットワーク制御部を
探索する
S1111
正常に動作している
ネットワーク制御部へハートビート
メッセージを送信
S1112
ネットワーク制御部から
の応答を一定時間待つ
S1113
S1114
応答あり
No
Yes
S1115
リトライカウンタ←0
S1117
リトライカウンタ+1
S1118
リトライカウンタ＞
リトライオーバ閾値
No
Yes
障害監視部へ障害通知
S1119
障害監視部からの応答待ち
S1120
フェイルオーバ処理開始
S1121
S1115
リトライカウンタ←0
一定時間待ち
S1116

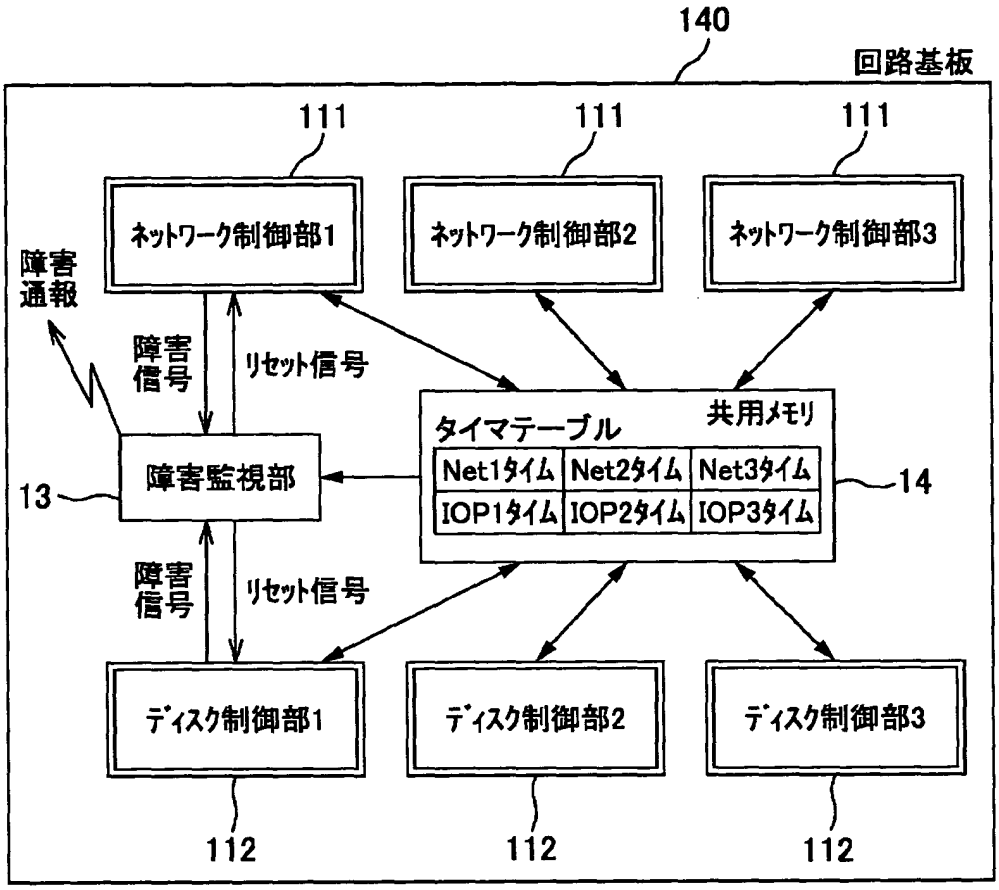
【図１２】


メッセージ受信
S1211
S1212
ハートビート
メッセージ？
No
Yes
S1214
障害監視部からの
動作していない旨の
通知か？
No
Yes
S1213
ハートビートメッセージをネットワーク
制御部へ送信する
S1215
障害のあったネットワーク制御部
への通信を抑止する
S1216
フェイルオーバ処理を開始
S1217
メッセージ破棄

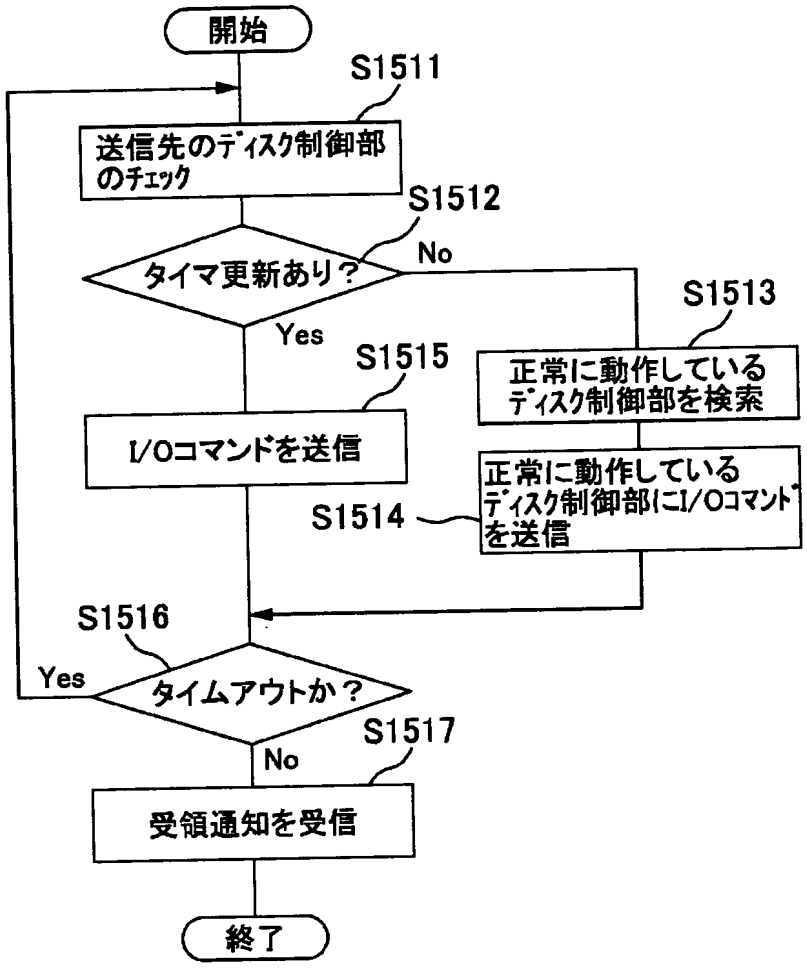
【図１３】


S1311
障害通知あり？
No
Yes
S1312
ハードウェア障害
No
Yes
S1315
ハートビートメッセージが送られてきていない旨の通知か？
No
Yes
S1316
そのネットワーク制御部の動作を停止させる
制御用メモリにステータス情報を書き込む
S1313
ネットワーク制御部へメッセージを送信
S1314

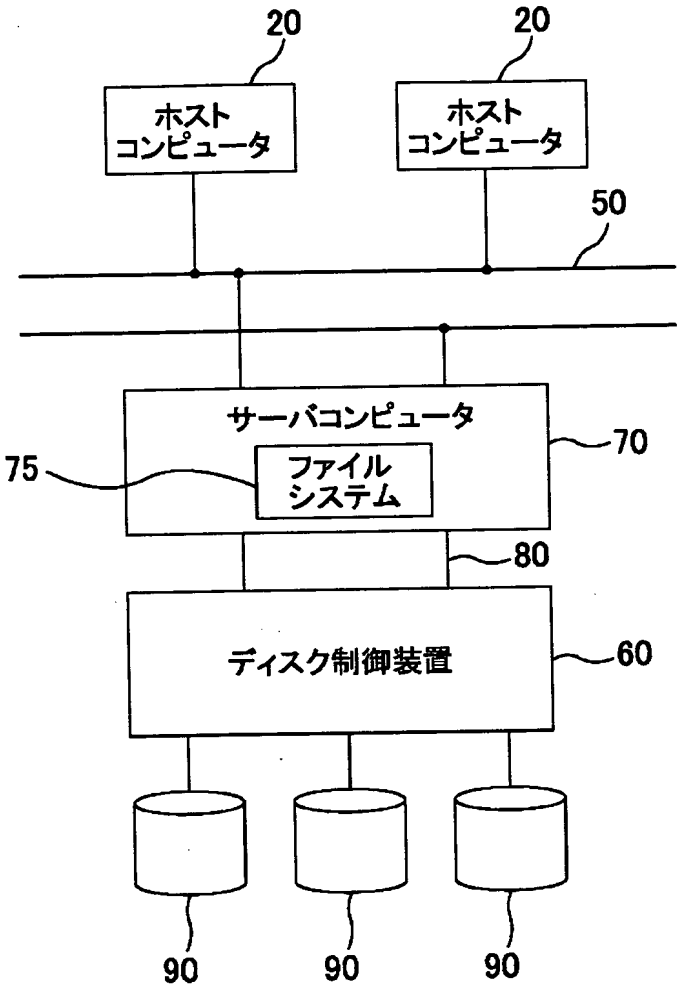
【図１４】


140
回路基板
111
111
111
ネットワーク制御部1
ネットワーク制御部2
ネットワーク制御部3
障害通報
障害信号
リセット信号
タイマテーブル
共用メモリ
Net1タイム
Net2タイム
Net3タイム
IOP1タイム
IOP2タイム
IOP3タイム
14
障害監視部
13
障害信号
リセット信号
ディスク制御部1
ディスク制御部2
ディスク制御部3
112
112
112

【図１５】


開始
S1511
送信先のディスク制御部のチェック
S1512
タイマ更新あり？
No
Yes
S1513
正常に動作しているディスク制御部を検索
S1515
I/Oコマンドを送信
正常に動作しているディスク制御部にI/Oコマンドを送信
S1514
S1516
Yes
タイムアウトか？
S1517
No
受領通知を受信
終了

【図１６】


20
20
ホスト
コンピュータ
ホスト
コンピュータ
50
サーバコンピュータ
70
75
ファイル
システム
80
ディスク制御装置
60
90
90
90

【書類名】　　要約書

【要約】

【解決手段】　ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うように構成したディスク制御装置を提供する。

【選択図】　　図１

出 願 人 履 歴 情 報

識別番号　〔０００００５１０８〕

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所